VISIT…

LANZAROTE Caliente.COM

# Глава 1. ЧИСЛЕННОЕ РЕШЕНИЕ ЛИНЕЙНЫХ АЛГЕБРАИЧЕСКИХ СИСТЕМ (СЛАУ)

В этой главе рассматривается одна из самых важных задач линейной алгебры – решение систем линейных алгебраических уравнений, в которых число уравнений равно числу неизвестных:

$$\begin{array}{l} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = f_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = f_2 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n = f_n \end{array} \qquad (1)$$

или в сокращенной записи:

$$\sum_{j=1}^{n} a_{ij}x_j = f_i,\ i = 1,2,\ldots,n.$$

Коэффициенты $a_{i,j}$ при неизвестных $x_j$ образуют матрицу системы (1)

$$A = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix}. \qquad (2)$$

Всюду на протяжении этой главы мы будем считать определитель матрицы отличным от нуля

$$\Delta = \det A \neq 0. \qquad (3)$$

В этом случае система (1) называется невырожденной. Решение невырожденной системы всегда существует и является единственным. Обсудим методы фактического построения этого решения.

## *§1. Прямые методы решения СЛАУ.*

Прямыми называются методы, которые позволяют получить точное решение невырожденной системы (1) за конечное число операций.

### 1.1. Формулы Крамера

Формулы Крамера представляют компоненты $x_j$ решения системы (1) в виде отношения двух определителей:

$$x_j = \Delta_j / \Delta,\ j = 1,2,\ldots,n, \qquad (4)$$

где

$$\Delta_j = \det A_j,\ j = 1,2,\ldots,n. \qquad (5)$$

Здесь матрица $A_j$ получается из матрацы $A$ заменой ее $j$-го столбца столбцом правых частей системы (1)

$$A_j = \begin{bmatrix} a_{11} & \dots & a_{1,j-1} & f_1 & a_{1,j+1} & \dots & a_{1n} \\ a_{21} & \dots & a_{2,j-1} & f_2 & a_{2,j+1} & \dots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{n1} & \dots & a_{n,j-1} & f_n & a_{n,j+1} & \dots & a_{nn} \end{bmatrix} \qquad (6)$$

С теоретической точки зрения формулы Крамера (4) дают исчерпывающее решение проблемы. Чтобы найти решение системы (1), нужно подсчитать $n+1$ определитель. Это можно сделать за конечное число арифметических операций. Однако с точки зрения практики важное значение имеет фактическое число необходимых операций. Здесь нас и поджидает главная трудность. Определитель $n$-ого порядка – это $n!$ слагаемых, каждое из которых является произведением $n$ чисел. Таким образом, для его вычисления нужно выполнить $(n-1)n!$ умножений и $n!$ сложений – всего $Q_n = n \cdot n!$ арифметических операций. Оценим это число. При $n \gg 1$ число $n!$ можно подсчитать с помощью асимптотической формулы Стирлинга:

$$n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n, \text{ так что } Q_n \approx \sqrt{2\pi} \cdot n^{\frac{3}{2}}\left(\frac{n}{e}\right)^n.$$

При умеренном значении $n = 20$ эта формула дает астрономическое число:

$$Q_{20} \approx 5 \cdot 10^{19}.$$

Компьютеру, производительность которого составляет $m$ операций/сек, для вычисления определителя двадцатого порядка понадобится время

$$T_{20} \approx (5 \cdot 10^{19} / m) \text{ сек.}$$

В частности, при $m = 10^{10}$ операций/сек получим

$$T_{20} \approx 5 \cdot 10^9 \text{ сек.} \approx 170 \text{ лет.}$$

Даже увеличение производительности компьютера на два, три порядка не спасает положения.

Такие результаты получены при $n = 20$, в то время, как в современных прикладных задачах приходится решать системы с $n = 10^6$ и более уравнений. Из проведенного анализа ясно, что рассчитывать решение СЛАУ по формулам Крамера с вычислением определителей «в лоб» невозможно, т. е. практическая ценность этих формул невелика.

### 1.2. Метод Гаусса.

Блестящий конструктивный выход из критической ситуации, описанной выше, дает метод Гаусса. Этот метод удобно условно разделить на два этапа. На первом этапе (прямой ход) система (1) приводится к треугольному виду. Затем на втором этапе (обратный ход) осуществляется последовательное отыскание неизвестных $x_1, \dots, x_n$ из этой треугольной системы.

Перейдем к подробному описанию метода Гаусса. Не ограничивая общности, будем считать, что коэффициент $a_{11}$, который называют ведущим элементом первого шага, отличен от нуля (в случае $a_{11} = 0$ поменяем местами уравнения с номерами 1 и $i$,

при котором $a_{i1} \neq 0$; поскольку система предполагается невырожденной, то такой номер $i$ заведомо найдется).

Разделим все члены первого уравнения на $a_{11}$ и введем в качестве новых коэффициентов $c_{1i}, i = 2,\ldots n$ и правой части $y_1$ отношения

$$c_{12} = \frac{a_{12}}{a_{11}}, \quad c_{13} = \frac{a_{13}}{a_{11}}, \quad \ldots \quad c_{1n} = \frac{a_{1n}}{a_{11}}, \ y_1 = \frac{f_1}{a_{11}}. \tag{7}$$

Вычтем из каждого $i$-го уравнения системы ($i = 2,\ldots,n$) первое уравнение умноженное на $a_{i1}$. Проделав это, мы исключим неизвестное $x_1$ из всех уравнений, кроме первого. Преобразованная таким образом система (1) примет эквивалентный вид:

$$\begin{aligned} x_1 + c_{12}x_2 + c_{13}x_3 + \ldots + c_{1n}x_n &= y_1 \\ a_{22}^{(1)}x_2 + a_{23}^{(1)}x_3 + \ldots + a_{2n}^{(1)}x_n &= f_2^{(1)} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ a_{n2}^{(1)}x_2 + a_{n3}^{(1)}x_3 + \ldots + a_{nn}^{(1)}x_n &= f_n^{(1)} \end{aligned}. \tag{8}$$

Значения новых коэффициентов и правых частей системы (8) вычисляются по формулам:

$$a_{ij}^{(1)} = a_{ij} - a_{i1}\frac{a_{1j}}{a_{11}}, \ f_i^{(1)} = f_i - a_{i1}\frac{f_1}{a_{11}}. \tag{9}$$

Естественно выделить из (8) «укороченную» систему, содержащую $n-1$ уравнение

$$\begin{aligned} a_{22}^{(1)}x_2 + a_{23}^{(1)}x_3 + \ldots + a_{2n}^{(1)}x_n &= f_2^{(1)} \\ a_{32}^{(1)}x_2 + a_{33}^{(1)}x_3 + \ldots + a_{3n}^{(1)}x_n &= f_3^{(1)} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ a_{n2}^{(1)}x_2 + a_{n3}^{(1)}x_3 + \ldots + a_{nn}^{(1)}x_n &= f_n^{(1)}. \end{aligned}$$

Продолжая далее процесс исключения, после $(n-1)$ шага редуцируем исходную систему к виду:

$$\begin{aligned} x_1 + c_{12}x_2 + c_{13}x_3 + \ldots + c_{1n}x_n &= y_1 \\ x_2 + c_{23}x_3 + \ldots + c_{2n}x_n &= y_2 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ x_{n-1} + c_{n-1,n}x_n &= y_{n-1} \\ x_n &= y_n \end{aligned} \tag{10}$$

или в матричной форме

$$C\mathbf{x} = \mathbf{y},$$

где матрица $C$ является верхней треугольной матрицей с единицами на главной диагонали

$$C = \begin{bmatrix} 1 & c_{12} & c_{13} & \dots & c_{1n} \\ 0 & 1 & c_{23} & \dots & c_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}. \quad (11)$$

Построение системы (10) завершает прямой ход метода Гаусса.

Обратный ход состоит в последовательном определении неизвестных из системы (10) в обратном порядке:

$$\begin{aligned} x_n &= y_n \\ x_{n-1} &= y_{n-1} - c_{n-1,n} x_n \\ x_{n-2} &= y_{n-2} - c_{n-2,n-1} x_{n-1} - c_{n-2,n} x_n \\ &\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ x_1 &= y_1 - c_{12} x_2 - c_{13} x_3 - \dots - c_{1n} x_n \end{aligned} \quad (12)$$

Подсчитаем число арифметических операций, которое требуется выполнить при решении СЛАУ по методу Гаусса. Первый шаг прямого хода, согласно формулам (7) и (9), требует $n$ делений и $n(n-1)$ сложений и умножений. Мы учитываем деления отдельно, поскольку для компьютера, как и для человека, это более сложная операция. Переходя последовательно от $n$ к $(n-1)$, потом от $(n-1)$ к $(n-2)$ и т.д. подсчитаем общее число арифметических операций на стадии прямого хода. Оно включает делений

$$Q_1 = n + (n-1) + \dots + 1 = \frac{1}{2} n (n+1), \quad (13)$$

сложений и умножений

$$Q_2 = n(n-1) + (n-1)(n-2) + \dots + 2 \cdot 1 = \frac{1}{3} n (n^2 - 1). \quad (14)$$

Обратный ход, согласно формулам (12), вообще не требует деления, а необходимое число сложений и умножений подсчитывается по формуле

$$Q_3 = 1 + 2 + \dots + (n-1) = \frac{1}{2} n (n-1). \quad (15)$$

Сравнивая (13) и (14) с (15), мы видим, что обратный ход существенно проще прямого. Сумма (14) и (15) дает общее число сложений и умножений, необходимое для решения СЛАУ по методу Гаусса:

$$Q = Q_2 + Q_3 = \frac{1}{3} n (n-1) \left( n + \frac{5}{2} \right) = \frac{1}{3} n^3 + O(n^2). \quad (16)$$

Оно не идет ни в какое сравнение с числом $n \cdot n!$, которое требуют формулы Крамера при прямом вычислении определителей.

Описанная выше процедура решения системы (1) методом Гаусса может оказаться неустойчивой по отношению к случайным ошибкам, которые неизбежны при компьютерных расчетах в результате округления чисел из-за конечной длины машинного слова. Действительно, предположим, что в процессе приведения системы (1) к треугольному виду (10) у матрицы $C$ (11) образовались большие по модулю

элементы: $\left|c_{i,j}\right| > 1$ и даже $\left|c_{i,j}\right| \gg 1$. Тогда при вычислении неизвестных по формулам (12) во время обратного хода умножение найденных с ошибками округления чисел $x_i$ на большие по модулю элементы матрицы $C$ приведет к увеличению этих ошибок. Наоборот, если матрица $C$ оказалась такой, что все ее элементы удовлетворяют условию

$$\left|c_{i,j}\right| \le 1, \qquad (17)$$

то роль ошибок округления в процессе вычислений будет нивелироваться.

Опишем, как можно добиться выполнения условия (17). Приступая к первому шагу прямого хода метода Гаусса, рассмотрим элементы $a_{1,j}$ первой строки матрицы $A$ и найдем среди них элемент наибольший по модулю. Пусть он имеет номер $j_1$. Поменяем в системе (1) первый столбец и столбец с номером $j_1$ местами, изменив соответствующим образом нумерацию неизвестных. В результате такой процедуры наибольший по модулю элемент первой строки станет ведущим элементом первого шага $a_{1,1}$. Благодаря этому элементы $c_{1,j}$ первой строки матрицы $C$, которые рассчитываются по формулам (7), будут удовлетворять неравенству (17).

Процедуру выделения наибольшего по модулю элемента в очередной строке и превращения его в ведущий элемент нужно затем повторять во время каждого шага прямого хода метода Гаусса. В этом случае все элементы $c_{i,j}$ треугольной матрицы $C$ (11) будут удовлетворять неравенствам (17), обеспечивая устойчивость метода по отношению к ошибкам округления, Такой способ коррекции называется выбором ведущего элемента по строке.

Поясним важность специального выбора ведущего элемента в каждой строке во время прямого хода метода Гаусса на простом примере. Рассмотрим систему трех уравнений с тремя неизвестными

$$\begin{aligned} 1.2357x_1 + 2.1742x_2 - 5.4834x_3 &= -2.0735 \\ 3.4873x_1 + 6.1365x_2 - 4.7483x_3 &= 4.8755 \\ 6.0696x_1 - 6.2163x_2 - 4.6921x_3 &= -4.8388. \end{aligned} \qquad (18)$$

Легко проверить, что ее решение имеет вид

$$x_1 = x_2 = x_3 = 1. \qquad (19)$$

Решим систему (18) с помощью метода Гаусса, не обращая внимание на величины элементов матрицы. Все результаты расчетов условимся представлять в виде чисел с плавающей запятой с пятью значащими цифрами. Тогда после прямого хода получим систему треугольного вида:

$$\begin{aligned} x_1 + 1.7595x_2 - 4.4375x_3 &= -1.6780 \\ x_2 + 15324x_3 &= 15324 \\ x_3 &= 0.99992. \end{aligned} \qquad (20)$$

Значение $x_3 = 0.99992$ выглядит вполне приемлемым. Однако для двух других неизвестных мы получим следующие значения: $x_2 = 2$, $x_1 = -0.75990$. Причина случившегося заключается в потере точности при вычислении $x_2$ из-за больших

значений коэффициента $c_{23}$ и правой части $y_2$ треугольной системы (20), которые вычислены с ошибками вследствие отбрасывания "лишних" значащих цифр.

Теперь воспользуемся процедурой выбора главного элемента по строке. Для этого в данном случае достаточно поменять местами первый и третий столбцы матрицы системы. В результате система примет вид:

$$\begin{aligned} -5.4834x_3 + 2.1742x_2 + 1.2357x_1 &= -2.0735 \\ -4.7483x_3 + 6.1365x_2 + 3.4873x_1 &= 4.8755 \\ -4.6921x_3 - 6.2163x_2 + 6.0696x_1 &= -4.8388 \end{aligned} \quad . \qquad (21)$$

При такой ее записи ведущим элементом первого шага становится число $-5.4834$. Оно является наибольшим по модулю элементом первой строки системы (21). Теперь применение метода Гаусса приводит к следующей системе с треугольной матрицей:

$$\begin{aligned} x_3 - 0.39651x_2 - 0.22535x_1 &= 0.37814 \\ x_2 + 0.56827x_1 &= 1.5682 \\ x_1 &= 0.99995 \end{aligned} \quad . \qquad (22)$$

Все ее элементы удовлетворяют неравенству (17). Осуществляя обратный ход, получим решение системы:

$$x_1 = 0.99995\,, x_2 = 0.99996\,,\ x_3 = 0.99999\,. \qquad (23)$$

Полученные значения неизвестных $x_i$ хорошо согласуются с ответом (19) в рамках принятой точности вычислений.

Нетрудно предвидеть, что при бесконтрольном применении метода Гаусса для решения больших систем ($n >> 1$) возможностей для потери точности становится еще больше, в то время как выполнение процедуры выбора ведущих элементов по строкам снимает эту проблему.

В заключение отметим, что первый этап метода Гаусса может быть использован для вычисления определителя матрицы $A$. Прямой ход метода Гаусса основан на многократном выполнении операции сложения одной из строк матрицы с другой строкой, взятой с некоторым множителем, что не меняет определителя. Следует лишь учесть, что при делении ведущей строки на ее диагональный элемент определитель также делится на этот элемент. Кроме того, иногда приходится переставлять столбцы при выборе главного элемента по строке. Поскольку определитель приведенной треугольной системы (матрицы $C$) всегда равен единице, то определитель $\Delta$ исходной системы равен

$$\Delta = \det A = (-1)^k\, a_{11} a_{22}^{(1)} \ldots a_{nn}^{(n-1)},$$

где $k$ – число перестановок столбцов в процессе редукции матрицы $A$ к треугольной матрице $C$.

### 1.3. Системы с диагональным преобладанием.

**Определение.**

*Назовем систему (1) системой с диагональным преобладанием по строке, если элементы матрицы $A$ (2) удовлетворяют неравенствам:*

$$\left|a_{i,i}\right| > \sum_{\substack{j=1 \\ j\neq i}}^{n} \left|a_{i,j}\right|,\ 1 \le i \le n \tag{24}$$

Неравенства (24) означают, что в каждой строке матрицы $A$ диагональный элемент выделен: его модуль больше суммы модулей всех остальных элементов той же строки.

**Теорема**

*Система с диагональным преобладанием всегда разрешима и притом единственным образом.*

Рассмотрим соответствующую однородную систему:

$$\sum_{j=1}^{n} a_{i,j} x_j = 0,\ 1 \le j \le n \tag{25}$$

Предположим, что она имеет нетривиальное решение $\overline{x}_j$, Пусть наибольшая по модулю компонента этого решения соответствует индексу $j = k$, т. е.

$$\left|\overline{x}_k\right| > 0,\ \left|\overline{x}_k\right| \ge \left|\overline{x}_j\right|,\ 1 \le j \le n. \tag{26}$$

Запишем $k$-ое уравнение системы (25) в виде

$$a_{k,k}\overline{x}_k = -\sum_{\substack{j=1 \\ j\neq k}}^{n} a_{k,j}\overline{x}_j$$

и возьмем модуль от обеих частей этого равенства. В результате получим:

$$\left|a_{k,k}\right|\left|\overline{x}_k\right| \le \sum_{\substack{j=1 \\ j\neq k}}^{n} \left|a_{k,j}\right|\left|\overline{x}_j\right| \le \left|\overline{x}_k\right| \sum_{\substack{j=1 \\ j\neq k}}^{n} \left|a_{k,j}\right|. \tag{27}$$

Сокращая неравенство (27) на множитель $\left|\overline{x}_k\right|$, который, согласно (26), не равен нулю, придем к противоречию с неравенством (24), выражающим диагональное преобладание. Полученное противоречие позволяет последовательно высказать три утверждения:

1. Однородная система (25) с диагональным преобладанием имеет только тривиальное решение.
2. Определитель матрицы $A$ с диагональным преобладанием не равен нулю.
3. Неоднородная система (1) с диагональным преобладанием всегда разрешима и притом единственным образом.

Последнее из них означает, что доказательство теоремы завершено.

### 1.4. Системы с трехдиагональной матрицей. Метод прогонки.

При решении многих задач приходится иметь дело с системами линейных уравнений вида:

$$A_i x_{i-1} + C_i x_i + B_i x_{i+1} = F_i,\ i = 1,...,n-1, \tag{28}$$

$$x_0 = q_0,\ x_n = q_n, \tag{29}$$

где коэффициенты $A_i, C_i, B_i$, правые части $F_i$ $(i = 1, \ldots, n-1)$ известны вместе с числами $q_0$ и $q_n$. Дополнительные соотношения (29) часто называют краевыми условиями для системы (28). Во многих случаях они могут иметь более сложный вид. Например:

$$x_0 = p_0 x_1 + q_0;\ x_n = p_n x_{n-1} + q_n,$$

где $p_0, q_0, p_n, q_n$ - заданные числа. Однако, чтобы не усложнять изложение, мы ограничимся простейшей формой дополнительных условий (29).

Пользуясь тем, что значения $x_0$ и $x_n$ заданы, перепишем систему (28) в виде:

$$\begin{aligned} &C_1 x_1 + B_1 x_2 && = F_1 - A_1 q_0 \\ &A_2 x_1 + C_2 x_2 + B_1 x_3 && = F_2 \\ &\vdots \\ &A_{n-1} x_{n-2} + C_{n-1} x_{n-1} && = F_{n-1} - B_{n-1} q_n \end{aligned} \qquad (30)$$

Матрица этой системы имеет трёхдиагональную структуру:

$$\begin{bmatrix} C_1 & B_1 & 0 & 0 & \cdots & 0 & 0 \\ A_2 & C_2 & B_2 & 0 & \cdots & 0 & 0 \\ 0 & A_3 & C_3 & B_3 & \cdots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & 0 & \ldots & A_{n-1} & C_{n-1} \end{bmatrix} \qquad (31)$$

Это существенно упрощает решение системы (28) благодаря специальному методу, получившему название метода прогонки.

Метод основан на предположении, что искомые неизвестные $x_i$ и $x_{i+1}$ связаны рекуррентным соотношением

$$x_i = \alpha_{i+1} x_{i+1} + \beta_{i+1},\ 0 \le i \le n-1. \qquad (32)$$

Здесь величины $\alpha_{i+1}$, $\beta_{i+1}$, получившие название прогоночных коэффициентов, подлежат определению, исходя из условий задачи (28), (29). Фактически такая процедура означает замену прямого определения неизвестных $x_i$ задачей определения прогоночных коэффициентов с последующим расчетом по ним величин $x_i$.

Для реализации описанной программы выразим с помощью соотношения (32) $x_{i-1}$ через $x_{i+1}$:

$$x_{i-1} = \alpha_i x_i + \beta_i = \alpha_i \alpha_{i+1} x_{i+1} + \alpha_i \beta_{i+1} + \beta_i$$

и подставим $x_{i-1}$ и $x_i$, выраженные через $x_{i+1}$, в исходные уравнения (28). В результате получим:

$$(A_i \alpha_i \alpha_{i+1} + C_i \alpha_{i+1} + B_i) x_{i+1} + A_i \alpha_i \beta_{i+1} + A_i \beta_i + C_i \beta_{i+1} - F_i = 0,$$
$$i = 1, 2, \ldots n-1.$$

Последние соотношения будут заведомо выполняться и притом независимо от решения, если потребовать, чтобы при $i = 1, 2, \ldots n-1$ имели место равенства:

$$A_i \alpha_i \alpha_{i+1} + C_i \alpha_{i+1} + B_i = 0,$$
$$A_i \alpha_i \beta_{i+1} + A_i \beta_i + C_i \beta_{i+1} - F_i = 0.$$

Отсюда следуют рекуррентные соотношения для прогоночных коэффициентов:

$$\alpha_{i+1} = \frac{-B_i}{A_i\alpha_i + C_i},\ \beta_{i+1} = \frac{F_i - A_i\beta_i}{A_i\alpha_i + C_i},\ i = 1,2,\dots n-1. \qquad (33)$$

Левое граничное условие $x_0 = q_0$ и соотношение $x_0 = \alpha_1 x_1 + \beta_1$ непротиворечивы, если положить

$$\alpha_1 = 0, \quad \beta_1 = q_0. \qquad (34)$$

Остальные значения коэффициентов прогонки $\alpha_2,\dots,\alpha_n$ и $\beta_2,\dots,\beta_n$ находим из (33), чем и завершаем этап вычисления прогоночных коэффициентов.

Далее, согласно правому граничному условию

$$x_n = q_n. \qquad (35)$$

Отсюда можно найти остальные неизвестные $x_{n-1},\dots x_1$ в процессе обратной прогонки с помощью рекуррентной формулы (32).

Число операций, которое требуется для решения системы общего вида (1) методом Гаусса, растет при увеличении $n$ пропорционально $n^3$. Метод прогонки сводится к двум циклам: сначала по формулам (33) рассчитываются прогоночные коэффициенты, затем с их помощью по рекуррентным формулам (32) находятся компоненты решения системы $x_i$. Это означает, что с увеличением размеров системы число арифметических операций будет расти пропорционально $n$, а не $n^3$. Таким образом, метод прогонки в пределах сферы своего возможного применения является существенно более экономичным. К этому следует добавить особую простоту его программной реализации на компьютере.

Во многих прикладных задачах, которые приводят к СЛАУ с трехдиагональной матрицей, ее коэффициенты удовлетворяют неравенствам:

$$|C_i| > |A_i| + |B_i|, \qquad (36)$$

которые выражают свойство диагонального преобладания. В частности, мы встретим такие системы в третьей и пятой главе.

Согласно теореме предыдущего раздела решение таких систем всегда существует и является единственным. Для них также справедливо утверждение, которое имеет важное значение для фактического расчета решения с помощью метода прогонки.

**Лемма**

*Если для системы с трехдиагональной матрицей выполняется условие диагонального преобладания (36), то прогоночные коэффициенты удовлетворяют неравенствам:*

$$|\alpha_i| \le 1. \qquad (37)$$

Доказательство проведем по индукции. Согласно (34) $\alpha_1 = 0$ , т. е. при $i = 1$ утверждение леммы верно. Допустим теперь, что оно верно для $\alpha_i$ и рассмотрим $\alpha_{i+1}$:

$$|\alpha_{i+1}| = \left|\frac{B_i}{C_i + A_i\alpha_i}\right| \le \frac{|B_i|}{|C_i| - |A_i|} \le 1. \qquad (38)$$

Итак, индукция от $i$ к $i+1$ обоснована, что и завершает доказательство леммы.

Неравенство (37) для прогоночных коэффициентов $\alpha_i$ делает прогонку устойчивой. Действительно, предположим, что компонента решения $x_i$ в результате процедуры округления рассчитана с некоторой ошибкой. Тогда при вычислении

следующей компоненты $x_{i-1}$ по рекуррентной формуле (32) эта ошибка, благодаря неравенству (37), не будет нарастать.

## §2. Обусловленность СЛАУ.

Серьезным препятствием при решении систем линейных алгебраических уравнений может оказаться возможность заметного отклонения приближенного решения от точного из-за незначительных возмущений правых частей уравнений, которые неизбежно возникают в приближенных вычислениях. Причиной такого нежелательного эффекта часто оказывается так называемая плохая обусловленность матрицы системы линейных уравнений.

### 2.1. Норма матрицы.

Рассмотрим линейное вещественное евклидово пространство $E_n$, элементами которого являются вектора в виде упорядоченной системы $n$ чисел $\mathbf{x} = \{x_1, \ldots x_n\}$. В пространстве $E_n$ определены скалярное произведение

$$(\mathbf{x}, \mathbf{y}) = x_1 y_1 + \ldots + x_n y_n \tag{39}$$

и евклидова норма

$$\|\mathbf{x}\| = \sqrt{(\mathbf{x}, \mathbf{x})} = \sqrt{x_1^2 + \ldots + x_n^2}\,, \tag{40}$$

удовлетворяющая трем аксиомам нормы:

1. $\|\mathbf{x}\| \ge 0$, $\|\mathbf{x}\| = 0$ тогда и только тогда, когда $\mathbf{x} = 0$;
2. $\|\alpha\mathbf{x}\| = |\alpha| \; \|\mathbf{x}\| \quad \forall \alpha, \mathbf{x}$;
3. $\|\mathbf{x} + \mathbf{y}\| \le \|\mathbf{x}\| + \|\mathbf{y}\|$ (неравенство треугольника).

Для скалярного произведения справедливо неравенство Коши-Буняковского $|(\mathbf{x}, \mathbf{y})| \le \|\mathbf{x}\| \; \|\mathbf{y}\|$.

Рассмотрим квадратную матрицу $A$ размером $n \times n$. Она определяет в пространстве $E_n$ линейное преобразование

$$\mathbf{y} = A\mathbf{x} \tag{41}$$

или

$$y_i = \sum_{j=1}^{n} a_{ij} x_j\,, \quad i = 1, \ldots, n\,.$$

Введем величину

$$\|A\| = \sup_{x \ne 0} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|}\,, \tag{42}$$

которую принято называть нормой матрицы $A$, согласованной с нормой вектора $\|\mathbf{x}\|$. Записывая ненулевой вектор $\mathbf{x}$ в виде

$$\mathbf{x} = \|\mathbf{x}\|\mathbf{z}\,,$$

где $\mathbf{z}$ вектор единичной длины: $\|\mathbf{z}\| = 1$, получим представление для нормы, эквивалентное (42)

$$\|A\| = \sup_{\|\mathbf{z}\|=1} \|A\mathbf{z}\|\,. \tag{43}$$

Отсюда следует, что в конечномерном пространстве норма матрицы ограничена, причем на единичной сфере всегда найдется такой вектор $\mathbf{z}_0$, что

$$\|A\| = \|A\mathbf{z}_0\|.$$

Наконец, из определения нормы (42) следует, что

$$\|A\mathbf{x}\| \le \|A\| \cdot \|\mathbf{x}\|. \quad (44)$$

Это простое неравенство лежит в основе всех дальнейших оценок.

## 2.2. Корректность решения СЛАУ.

Следуя Адамару, будем называть математическую задачу корректной, если выполняются три условия:

1. Решение задачи существует.
2. Решение задачи единственное.
3. Решение задачи непрерывно зависит от входных данных.

Обсудим с точки зрения этого определения задачу решения СЛАУ с неравным нулю определителем

$$A\mathbf{x} = \mathbf{f}, \quad (45)$$

считая матрицу $A$ фиксированной и рассматривая в качестве входных данных вектор правых частей системы $\mathbf{f} = \{f_1, f_2, \ldots, f_n\} \in E_n$.

Условие $\Delta \ne 0$ гарантирует существование у матрицы $A$ обратной матрицы $A^{-1}$, через которую решение системы (45) можно записать в виде

$$\mathbf{x} = A^{-1}\mathbf{f}. \quad (46)$$

Пусть теперь правая часть подверглась возмущению $\delta\mathbf{f}$ и стала равной $\tilde{\mathbf{f}} = \mathbf{f} + \delta\mathbf{f}$. Тогда, согласно (46), решение $\tilde{\mathbf{x}}$ возмущенной системы

$$A\tilde{\mathbf{x}} = \tilde{\mathbf{f}} \quad (47)$$

тоже можно записать через обратную матрицу $A^{-1}$:

$$\tilde{\mathbf{x}} = A^{-1}\tilde{\mathbf{f}} = A^{-1}\mathbf{f} + A^{-1}\delta\mathbf{f} = \mathbf{x} + \delta\mathbf{x}, \quad (48)$$

где

$$\delta\mathbf{x} = A^{-1}\delta\mathbf{f}. \quad (49)$$

Отсюда получаем

$$\|\delta\mathbf{x}\| \le \|A^{-1}\| \|\delta\mathbf{f}\|. \quad (50)$$

Неравенство (50) доказывает непрерывную зависимость возмущения решения $\delta\mathbf{x}$ от возмущения правой части $\delta\mathbf{f}$:

$$\|\delta\mathbf{x}\| \to 0 \text{ при } \|\delta\mathbf{f}\| \to 0. \quad (51)$$

Это означает, что решение СЛАУ с неравным нулю определителем $\Delta$ - корректная математическая задача: для нее выполняются все три требования корректности Адамара.

## 2.3. Число обусловленности матрицы.

Исходное уравнение (45) позволяет написать неравенство:

$$\|\mathbf{f}\| \le \|A\| \|\mathbf{x}\|. \quad (52)$$

Перемножая его с неравенством того же знака (50), получим:

$$\|\mathbf{f}\|\|\delta\mathbf{x}\| \le \|A\|\|A^{-1}\|\|\mathbf{x}\|\|\delta\mathbf{f}\|. \tag{53}$$

Пусть $\mathbf{f} \neq 0$, тогда, согласно (46), $\mathbf{x} \neq 0$ и неравенство (53) можно переписать в виде:

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \le M_A \frac{\|\delta\mathbf{f}\|}{\|\mathbf{f}\|}, \tag{54}$$

где

$$M_A = \|A\| \cdot \|A^{-1}\|. \tag{55}$$

Число $M_A$ называется числом обусловленности матрицы $A$. Оно позволяет оценить относительную погрешность решения через относительную погрешность возмущения правой части. Поскольку исходная система (45) линейная, оценка относительной погрешности является более естественной, чем оценка абсолютной погрешности. Чем больше $M_A$, тем резче реагирует решение на возмущение правой части. Поэтому матрицы с большим числом обусловленности и соответствующие им СЛАУ называют плохо обусловленными. Для оценки роли, которую играет число обусловленности при решении линейных алгебраических систем, разберем задачу.

**Задача 1**

*Рассмотреть систему двух уравнений*

$$\begin{matrix} x_1 + 0 \cdot x_2 = 1 \\ x_1 + 0.01 \cdot x_2 = 1 \end{matrix}, \quad A = \begin{bmatrix} 1 & 0 \\ 1 & 0.01 \end{bmatrix}, \quad \mathbf{f} = \{1, 1\} \tag{56}$$

*и соответствующую ей возмущенную систему*

$$\begin{matrix} x_1 + 0 \cdot x_2 = 1 \\ x_1 + 0.01 \cdot x_2 = 1.01 \end{matrix}, \quad A = \begin{bmatrix} 1 & 0 \\ 1 & 0.01 \end{bmatrix}, \quad \tilde{\mathbf{f}} = \{1, 1.01\}. \tag{57}$$

*Выписать решения этих систем, подсчитать погрешность возмущения правой части и соответствующую ей погрешность возмущения решения. Найти число обусловленности матрицы $A$, составить с его помощью теоретическую оценку погрешности (54) и сравнить результат с результатом, полученным непосредственно по известным решениям систем.*

В данном случае определитель матрицы $A$ отличен от нуля

$$\Delta = \det A = 0.01,$$

т. е. обе системы невырожденные. Система (57) отличается от системы (56) возмущением правой части

$$\mathbf{f} = \{1, 1\}, \ \|\mathbf{f}\| = \sqrt{2}, \ \tilde{\mathbf{f}} = \{1, 1.01\}, \delta\mathbf{f} = \{0, 0.01\}, \ \|\delta\mathbf{f}\| = 0.01.$$

Решения систем (56) и (57) имеют вид:

$$\mathbf{x} = \{1, 0\}, \ \|\mathbf{x}\| = 1, \ \tilde{\mathbf{x}} = \{1, 1\}, \delta\mathbf{x} = \{0, 1\}, \ \|\delta\mathbf{x}\| = 1.$$

При этом

$$\frac{\|\delta\mathbf{f}\|}{\|\mathbf{f}\|} = \frac{0.01}{\sqrt{2}}, \ \frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} = 1. \tag{58}$$

Мы видим, что небольшое относительное возмущение правой части привело к сильному возмущению решения: относительная погрешность решения равна единице. Этот результат означает, что исходная система плохо обусловлена. Чтобы убедиться в

этом, подсчитаем число обусловленности матрицы $A$, напишем с его помощью теоретическую оценку (54) и сравним ее с фактическим результатом (58).

Выпишем линейное преобразование $\mathbf{y} = A\mathbf{x}$ отвечающее матрице системы

$$y_1 = x_1$$
$$y_2 = x_1 + 0.01x_2,$$

при этом

$$\|A\mathbf{x}\| = \sqrt{x_1^2 + \left(x_1 + 0.01x_2\right)^2} .$$

Наложим ограничение

$$x_1^2 + x_2^2 = 1,$$

тогда в силу (43)

$$\|A\mathbf{x}\| = \max\sqrt{2x_1^2 + 0.0001x_2^2 + 0.02x_1x_2} ,\ x_1^2 + x_2^2 = 1 .$$

Если положить $x_1 = \cos\varphi$, $x_2 = \sin\varphi$, то задача сведется к отысканию максимума выражения

$$g(\varphi) = \sqrt{2\cos^2\varphi + 0.02\sin\varphi\cos\varphi + 0.0001\sin^2\varphi} ,$$

зависящего только от одной переменной $\varphi$, $0 \le \varphi \le 2\pi$.

Переходя к тригонометрическим функциям двойного угла

$$2\cos^2\varphi = 1 + \cos 2\varphi,\ 2\sin^2\varphi = 1 - \cos 2\varphi,\ 2\sin\varphi\cos\varphi = \sin 2\varphi,$$

сведем подрадикальное выражение к виду:

$$1.00005 + 0.01\sin 2\varphi + 0.99995\cos 2\varphi$$

Для комбинации

$$B_1\cos 2\varphi + B_2\sin 2\varphi = \sqrt{B_1^{\ 2} + B_2^{\ 2}}\cos(2\varphi - \varphi_0),\ 0 \le \varphi \le 2\pi,$$

где

$$\varphi_0 = \mathrm{arctg}\left( B_1 / B_2 \right),\ B_1 = 0.99995,\ B_2 = 0.01,$$

максимальное значение равно

$$\sqrt{B_1^{\ 2} + B_2^{\ 2}} = \sqrt{0.99995^2 + 0.01^2} .$$

Следовательно

$$\|A\| = \sqrt{1.00005 + \sqrt{0.99995^2 + 0.01^2}} .$$

С приемлемой точностью это число равно $\sqrt{2}$: $\|A\| \approx \sqrt{2}$.

Аналогичным образом находится норма обратной матрицы

$$A^{-1} = \begin{bmatrix} 1 & 0 \\ -100 & 100 \end{bmatrix},\ \|A^{-1}\| \approx 100\sqrt{2} .$$

Таким образом, в данном примере

$$M_A = \|A\| \cdot \|A^{-1}\| \approx 200. \qquad (59)$$

В результате теоретическая оценка (54) принимает вид:

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \le 200\frac{\|\delta\mathbf{f}\|}{\|\mathbf{f}\|}$$

Она согласуется с результатом (58), который мы получили, непосредственно решая системы (56) и (57).

В процессе решения задачи мы убедились в том, что подсчет числа обусловленности является сложной задачей, особенно с учетом того, что нужно вычислять норму не только прямой, но и обратной матрицы. Поэтому желательно получить какие-нибудь конструктивные оценки этой важнейшей характеристики системы.

## 2.4. Оценка числа обусловленности.

Для числа обусловленности матрицы $A$ справедливо неравенство

$$M_A \geq |\lambda_{\max}| / |\lambda_{\min}|, \tag{60}$$

где $\lambda_{\min}$ и $\lambda_{\max}$ соответственно минимальное и максимальное по модулю значения характеристических чисел матрицы $A$. Соотношение (60) корректно, поскольку в силу невырожденности матрицы $\lambda_{\min} \neq 0$.

В самом деле пусть $\mathbf{y}$ - собственный вектор линейного преобразования, связанного с матрицей $A$, отвечающий $\lambda_{\max}$:

$$A\mathbf{y} = \lambda_{\max}\mathbf{y},$$

тогда

$$|\lambda_{\max}|\|\mathbf{y}\| = \|A\mathbf{y}\| \leq \|A\| \cdot \|\mathbf{y}\|,$$

и, следовательно, поскольку $\|\mathbf{y}\| \neq 0$

$$|\lambda_{\max}| \leq \|A\|.$$

Аналогичным образом для собственного вектора $\mathbf{z}$, связанного с $\lambda_{\min}$, имеем

$$A\mathbf{z} = \lambda_{\min}\mathbf{z}$$

или

$$A^{-1}\mathbf{z} = \frac{1}{\lambda_{\min}}\mathbf{z}.$$

Отсюда следует оценка

$$\frac{1}{|\lambda_{\min}|} \leq \|A^{-1}\|.$$

Перемножая два последних неравенства, придем к утверждению (60).

Если матрица симметричная $A = A^*$, то все её характеристические значения вещественны, причем

$$\|A\| = |\lambda_{\max}| \text{ и } \|A^{-1}\| = \frac{1}{|\lambda_{\min}|},$$

поэтому для таких матриц

$$M_A = \frac{|\lambda_{\max}|}{|\lambda_{\min}|}. \tag{61}$$

Из полученной оценки для $M_A$ следуют два важных вывода:

1) $M_A \geq 1$;

2) Число обусловленности тем больше, чем больше разброс характеристических чисел матрицы. Поэтому с увеличением размера матрицы, вообще говоря, её обусловленность имеет тенденцию к ухудшению.

Возвращаясь к рассмотренной выше задаче, без труда находим: $\lambda_{\min} = 0.01$, $\lambda_{\max} = 1$ и, следовательно, справедлива оценка снизу

$$M_A \geq \lambda_{\max} / \lambda_{\min} = 100,$$

причем точность этой оценки невысока, но порядок она передает правильно.

В заключение данного параграфа еще раз отметим, что для систем уравнений с большой размерностью "хорошая" обусловленность ($M_A \sim 1$) является скорее исключением, чем правилом и обычно приходится иметь дело с плохо обусловленными матрицами ($M_A \gg 1$), причем получение оценки числа обусловленности вызывает большие трудности.

## *§3. Итерационные методы.*

### 3.1. Построение итерационных последовательностей.

Мы видели, что процедура решения СЛАУ

$$A\mathbf{x} = \mathbf{f} \tag{62}$$

с плохо обусловленной матрицей $A$ может приводить к существенным отклонениям получаемого ответа от точного решения при незначительных возмущениях правой части. Однако появление таких возмущений неизбежно, например, при преобразовании вектора правых частей в методе Гаусса из-за ошибок округления при выполнении арифметических операций. Чем выше порядок матрицы, тем больше может оказаться результирующая погрешность.

Этого недостатка лишены итерационные методы решения СЛАУ. При их применении ответ получается в процессе построения последовательных приближений (итераций) $\mathbf{x}_k = \left\{x_1^k, x_2^k, \dots x_n^k\right\}$, сходящихся к решению системы (62) в пространстве $E_n$ с евклидовой нормой $\|x\|$

$$\lim_{k\to\infty} \mathbf{x}_k = \mathbf{x} \tag{63}$$

Здесь при записи вектора $\mathbf{x}_k$ через его компоненты $x_i^k$ нижний индекс $i$ означает номер компоненты ($1 \leq i \leq n$), верхний индекс $k$ - номер итерации. Сходимость последовательности $\mathbf{x}_k$ к решению системы $\mathbf{x}$ означает, что

$$\lim_{k\to\infty} \|\mathbf{x}_k - \mathbf{x}\| = \lim_{k\to\infty} \sqrt{\left(x_1^k - x_1\right)^2 + \left(x_2^k - x_2\right)^2 + \dots + \left(x_n^k - x_n\right)^2} = 0. \tag{64}$$

Необходимым и достаточным условием предельного равенства (64) в конечномерном евклидовом пространстве $E_n$ является покомпонентная сходимость:

$$\lim_{k\to\infty} x_i^k = x_i, \; 1 \leq i \leq n.$$

Сходимость обеспечивает принципиальную возможность получить в процессе итераций ответ с любой наперед заданной степенью точности.

С итерационными последовательностями вы встречались. Каждый следующий член такой последовательности выражается через предыдущие, уже известные. Если, например, формула для вычисления очередного члена последовательности имеет вид:

$$\mathbf{x}_{k+1} = F\left(\mathbf{x}_k, \mathbf{x}_{k-1}, \ldots, \mathbf{x}_{k-m+1}\right),$$

то говорят о $m$-шаговом итерационном алгоритме. В частности, в простейшем случае очередной член последовательности $\mathbf{x}_{k+1}$ может выражается только через предыдущий $\mathbf{x}_k$:

$$\mathbf{x}_{k+1} = F\left(\mathbf{x}_k\right).$$

Такие итерационные алгоритмы называют одношаговыми.

При обсуждении итерационных методов решения СЛАУ мы ограничимся линейными одношаговыми алгоритмами, которые обычно записывают в стандартной канонической форме:

$$B_{k+1}\frac{\mathbf{x}_{k+1} - \mathbf{x}_k}{\tau_{k+1}} + A\mathbf{x}_k = \mathbf{f}, \ \det B_{k+1} \neq 0, \ \tau_{k+1} > 0. \quad (65)$$

В такой записи процесс характеризуется последовательностью матриц $B_{k+1}$ и числовых параметров $\tau_{k+1}$, которые называют итерационными параметрами. Если матрицы $B_{k+1}$ и параметры $\tau_{k+1}$ не меняются в процессе итераций, т. е. не зависят от индекса $k$, то итерационный процесс называется стационарным.

Перепишем формулу (65) в виде

$$B_{k+1}\mathbf{x}_{k+1} = \mathbf{F}_{k+1}, \quad (66)$$

где

$$\mathbf{F}_{k+1} = \left(B_{k+1} - \tau_{k+1}A_{k+1}\right)\mathbf{x}_k + \tau_{k+1}\mathbf{f}. \quad (67)$$

Мы видим, что построение очередной итерации сводится к решению системы уравнений (66) с правой частью (67), зависящей от предыдущей итерации $\mathbf{x}_k$. Такую задачу приходится решать многократно, поэтому матрицы $B_{k+1}$ следует выбирать достаточно простыми. Если построение отдельных итераций будет соизмеримым по сложности с решением исходной задачи, то метод окажется лишенным практического смысла.

Наиболее прост в реализации итерационный процесс с единичной матрицей: $B_{k+1} = E$. В этом случае формулы (66), (67) дают явное выражение очередной итерации через предыдущую:

$$\mathbf{x}_{k+1} = \left(E - \tau_{k+1}A\right)\mathbf{x}_k + \tau_{k+1}\mathbf{f}. \quad (68)$$

Из неявных итерационных методов выделим сравнительно легко реализуемые методы с диагональными матрицами: $B_{k+1} = D_{k+1}$ и верхними или нижними треугольными матрицами: $B_{k+1} = T_{k+1}$.

### 3.2. Проблема сходимости итерационного процесса.

Итерационный процесс может быть использован для решения СЛАУ только при условии сходимости. Для исследования его сходимости введем две характеристики. Первая из них – погрешность решения:

$$\mathbf{z}_k = \mathbf{x}_k - \mathbf{x}. \quad (69)$$

Смысл этого вектора ясен. Сходимость итерационного процесса согласно (63) и (64) означает, что

$$\lim_{k\to\infty}\mathbf{z}_k = 0\,,\ \lim_{k\to\infty} z_i^k = 0\,,\ 1\le i\le n\,. \qquad (70)$$

Вторая характеристика – невязка:

$$\boldsymbol{\psi}_k = A\mathbf{x}_k - \mathbf{f}\,. \qquad (71)$$

Она показывает, насколько хорошо или, наоборот, плохо член итерационной последовательности $\mathbf{x}_k$ удовлетворяет исходной системе.

Установим связь между $\mathbf{z}_k$ и $\boldsymbol{\psi}_k$:

$$\boldsymbol{\psi}_k = A\mathbf{x}_k - \mathbf{f} = A(\mathbf{z}_k + \mathbf{x}) - \mathbf{f} = A\mathbf{z}_k\,. \qquad (72)$$

Можно также написать обратное соотношение:

$$\mathbf{z}_k = A^{-1}\boldsymbol{\psi}_k\,. \qquad (73)$$

Из формул (72) и (73) вытекают оценки:

$$\|\boldsymbol{\psi}_k\| \le \|A\|\cdot\|\mathbf{z}_k\|,\ \|\mathbf{z}_k\| \le \left\|A^{-1}\right\|\cdot\|\boldsymbol{\psi}_k\|. \qquad (74)$$

Они показывают, что погрешность решения $\mathbf{z}_k$ стремится к нулю тогда и только тогда, когда стремится к нулю невязка $\boldsymbol{\psi}_k$. Этот результат позволяет судить о сходимости или расходимости итерационного процесса по поведению невязки, которая доступна прямому вычислению и благодаря этому может контролироваться.

При исследовании сходимости итерационных методов большую роль играют свойства матриц $A$ и $B_{n+1}$, в первую очередь такие как самосопряженность и знакоопределенность. Напомним, что в вещественном евклидовом пространстве $E_n$ для каждого линейного преобразования существует единственное сопряженное к нему линейное преобразование, определяемое тождественным равенством скалярных произведений:

$$(A\mathbf{x},\mathbf{y}) = (\mathbf{x}, A^*\mathbf{y})\,,\ \forall \mathbf{x},\mathbf{y}\in E_n\,. \qquad (75)$$

В частности,

$$(A\mathbf{x},\mathbf{x}) = (\mathbf{x}, A^*\mathbf{x})\,,\ \forall \mathbf{x}\in E_n\,.$$

Преобразование называется самосопряженным, если

$$(A\mathbf{x},\mathbf{y}) = (\mathbf{x}, A\mathbf{y})\,,\ \forall \mathbf{x},\mathbf{y}\in E_n\,. \qquad (76)$$

Матрицы сопряженных преобразований в ортонормированном базисе связаны простым транспонированием:

$$a_{ij}^* = a_{ji}\,,\ \forall i,j = 1,\ldots,n\,.$$

Свойство самосопряженности преобразования равносильно в этом случае выполнению условия совпадения матриц $A$ и $A^*$:

$$a_{ij} = a_{ji} = a_{ij}^*\,,\ \forall i,j = 1,\ldots,n\,,$$

Как известно, любая матрица представима в виде:

$$A = \bar{A} + \tilde{A}\,, \qquad (77)$$

где

$$\bar{A} = \frac{A + A^*}{2} = \bar{A}^*,\ \tilde{A} = \frac{A - A^*}{2} = -\tilde{A}^*. \qquad (78)$$

Нетрудно видеть, что

$$\left(A\mathbf{x},\mathbf{x}\right)=\left(A^{*}\mathbf{x},\mathbf{x}\right)=\left(\bar{A}\mathbf{x},\mathbf{x}\right),$$
$$\left(\tilde{A}\mathbf{x},\mathbf{x}\right)=0. \tag{79}$$

В дальнейшем мы будем опираться на следующие важные свойства самосопряженных преобразований:

а) все собственные значения самосопряженного линейного преобразования (характеристические числа матрицы $A$) вещественны;

б) самосопряженное линейное преобразование всегда имеет полный набор линейно независимых собственных векторов, из которых можно образовать ортонормированный базис пространства $E_n$. В этом базисе матрица линейного преобразования принимает диагональный вид, причем на диагонали стоят все собственные значения этого преобразования с учетом их кратности.

Наконец, матрица линейного преобразования $A$ называется положительно определенной, если для любого, отличного от нуля $\mathbf{x}\in E_n$ :

$$\left(A\mathbf{x},\mathbf{x}\right)>0,\ \sum_{i,j=1}^{n}a_{ij}x_ix_j>0,\ \forall\mathbf{x}\in E_n,\mathbf{x}\neq 0. \tag{80}$$

Для краткости, если это не вызывает недоразумений, будем часто писать $A>0$.

Необходимым и достаточным условием положительной определенности самосопряженной матрицы $A$ является критерий Сильвестра, из которого в частности следует строгая положительность всех диагональных элементов:

$$a_{i,i}>0,\ 1\le i\le n. \tag{81}$$

Условимся обозначать собственные векторы линейного преобразования с матрицей $A$ как $\mathbf{e}_i$, её характеристические числа как $\lambda_i$, координаты произвольного вектора $\mathbf{x}$ в ортонормированном базисе из собственных векторов $\mathbf{e}_i$ как $\xi_i$.

Для дальнейшего рассмотрения будут полезны три леммы.

**Лемма 1.**

*Для того, чтобы симметричная* $(A=A^{*})$ *матрица была положительно определенной, необходимо и достаточно, чтобы все её характеристические числа были положительны:* $\lambda_i>0$.

*Необходимость.* Выберем любой собственный вектор $\mathbf{e}_i$ линейного преобразования с матрицей $A$, тогда

$$(A\mathbf{e}_i,\mathbf{e}_i)=\lambda_i>0.$$

*Достаточность.* Расположим для определенности все характеристические значения матрицы $A=A^{*}$ в порядке убывания:

$$\lambda_1\ge\lambda_2\ge\ldots\ge\lambda_n>0.$$

Поскольку по условию леммы $\lambda_i>0$, то в ортонормированном базисе из собственных векторов преобразования с матрицей $A$ для любого $\mathbf{x}\neq 0$ имеем

$$(A\mathbf{x},\mathbf{x})=\sum_{i=1}^{n}\lambda_i\xi_i^2>0,\ \forall\{\xi_i\},\ \left(\sum_{i=1}^{n}\xi_i^2>0\right).$$

Поэтому, очевидно, что $A>0$.

**Лемма 2.**

*Пусть* $A = A^* > 0$, *и* $\lambda_1 \geq \ldots \geq \lambda_n > 0$ *- упорядоченный набор характеристических чисел этой матрицы, тогда*

$$\lambda_n \|\mathbf{x}\|^2 \leq (A\mathbf{x}, \mathbf{x}) \leq \lambda_1 \|\mathbf{x}\|^2 . \qquad (82)$$

Доказательство предлагается провести самостоятельно.

**Лемма 3.**

*Если* $A > 0$, *то всегда найдется постоянное число* $\delta > 0$, *такое что*

$$(A\mathbf{x}, \mathbf{x}) \geq \delta \|\mathbf{x}\|^2, \ \forall \mathbf{x} \in E_n \qquad (83)$$

Доказательство.

Если $A = A^*$, то достаточно положить $\delta = \lambda_n$. В общем случае напомним, что согласно (79)

$$(A\mathbf{x}, \mathbf{x}) = (\overline{A}\mathbf{x}, \mathbf{x}) > 0,$$

где $\overline{A} = \overline{A}^*$, поэтому согласно предыдущей лемме

$$(A\mathbf{x}, \mathbf{x}) = (\overline{A}\mathbf{x}, \mathbf{x}) \geq \overline{\lambda}_n \|\mathbf{x}\|^2,$$

где $\overline{\lambda}_n > 0$ - минимальное характеристическое число матрицы $\overline{A} = (A + A^*)/2$. Полагая, что $\delta = \overline{\lambda}_n$, приходим к требуемому неравенству (83).

### 3.3. Достаточные условия сходимости итерационного процесса.

В этом разделе мы рассмотрим стационарный итерационный процесс (65), когда матрица $B$ и итерационный параметр $\tau$ не зависят от индекса $k$, и докажем следующую теорему о достаточных условиях его сходимости.

**Теорема Самарского**

*Пусть* $A$ *- самосопряженная положительно определенная матрица:*

$$A = A^*, \ A > 0, \qquad (84)$$

$B - \frac{\tau}{2}A$ *- положительно определенная матрица,* $\tau$ *- положительное число:*

$$B - \frac{\tau}{2}A > 0, \ \tau > 0. \qquad (85)$$

*Тогда при любом выборе нулевого приближения* $\mathbf{x}_0$ *итерационный процесс, который определяется рекуррентной формулой (65), сходится к решению исходной системы (62).*

Прежде, чем переходить к доказательству теоремы, обсудим более подробно главное ее требование – положительную определенность матрицы $B - \frac{\tau}{2}A$. Это требование можно переписать в виде:

$$\left(B\mathbf{x}, \mathbf{x}\right) > \frac{\tau}{2}\left(A\mathbf{x}, \mathbf{x}\right), \ \forall \mathbf{x} \in E_n, \ \mathbf{x} \neq 0. \qquad (86)$$

т. е. оно, в частности, предполагает, что матрица $B$ является положительно определенной. Кроме того, неравенство (86) определяет интервал, в котором может изменяться параметр $\tau$:

$$0<\tau<\tau_0=\inf_{x\neq 0}\frac{2(B\mathbf{x},\mathbf{x})}{(A\mathbf{x},\mathbf{x})}. \tag{87}$$

После этих замечаний перейдем к доказательству теоремы. Выразим из соотношения (69) $\mathbf{x}_k$ через $\mathbf{z}_k$:

$$\mathbf{x}_k=\mathbf{z}_k+\mathbf{x}$$

и подставим в рекуррентную формулу для итерационной последовательности (65). В результате получим:

$$B\frac{\mathbf{z}_{k+1}-\mathbf{z}_k}{\tau}+A\mathbf{z}_k=0. \tag{88}$$

Отличие итерационной формулы (88) от (65) заключается в том, что она является однородной.

Матрица $B$ - положительно определенная. Следовательно она невырожденная и имеет обратную $B^{-1}$. С ее помощью рекуррентное соотношение (88) можно разрешить относительно $\mathbf{z}_{k+1}$:

$$\mathbf{z}_{k+1}=\mathbf{z}_k-\tau B^{-1}A\mathbf{z}_k=\mathbf{z}_k-\tau\boldsymbol{\omega}_k, \tag{89}$$

где

$$\boldsymbol{\omega}_k=B^{-1}A\mathbf{z}_k, \text{ так что } A\mathbf{z}_k=B\boldsymbol{\omega}_k. \tag{90}$$

Умножая обе части равенства (89) слева на матрицу $A$, получим еще одно рекуррентное соотношение

$$A\mathbf{z}_{k+1}=A\mathbf{z}_k-\tau A\boldsymbol{\omega}_k. \tag{91}$$

Рассмотрим последовательность положительных функционалов:

$$J_k=(A\mathbf{z}_k,\mathbf{z}_k). \tag{92}$$

Составим аналогичное выражение для $J_{k+1}$ и преобразуем его с помощью рекуррентных формул (89) и (91):

$$\begin{aligned}J_{k+1}&=(A\mathbf{z}_k-\tau A\boldsymbol{\omega}_k,\mathbf{z}_k-\tau\boldsymbol{\omega}_k)=(A\mathbf{z}_k,\mathbf{z}_k)-\tau(A\boldsymbol{\omega}_k,\mathbf{z}_k)-\\&-\tau(A\mathbf{z}_k,\boldsymbol{\omega}_k)+\tau^2(A\boldsymbol{\omega}_k,\boldsymbol{\omega}_k).\end{aligned} \tag{93}$$

Из самосопряженности матрицы $A$ и формулы (90) следует

$$(A\boldsymbol{\omega}_k,\mathbf{z}_k)=(A\mathbf{z}_k,\boldsymbol{\omega}_k)=(B\boldsymbol{\omega}_k,\boldsymbol{\omega}_k).$$

В результате формула (93) принимает вид:

$$J_{k+1}=J_k-2\tau(B\boldsymbol{\omega}_k,\boldsymbol{\omega}_k)+\tau^2(A\boldsymbol{\omega}_k,\boldsymbol{\omega}_k)=J_k-2\tau\left(\left(B-\frac{\tau}{2}A\right)\boldsymbol{\omega}_k,\boldsymbol{\omega}_k\right). \tag{94}$$

Таким образом, последовательность функционалов $J_k$ с учетом условия $B-\frac{\tau}{2}A>0$ образует монотонно невозрастающую последовательность, ограниченную снизу нулем

$$J_k\geq J_{k+1}\geq\cdots\geq 0. \tag{95}$$

Поэтому она сходится. Далее, согласно лемме 3

$$\left(\left(B-\frac{\tau}{2}A\right)\boldsymbol{\omega}_k,\boldsymbol{\omega}_k\right)\geq\delta\|\boldsymbol{\omega}_k\|^2,$$

где $\delta > 0$ - строго положительная константа. В результате, согласно (94) и (95) будем иметь

$$J_{k+1} - J_k = 2\tau\left(\left(B - \frac{\tau}{2}A\right)\boldsymbol{\omega}_k, \boldsymbol{\omega}_k\right) \geq 2\tau\delta\left\|\boldsymbol{\omega}_k\right\|^2. \quad (96)$$

Из этого неравенства и сходимости последовательности функционалов $J_k$ следует, что $\left\|\boldsymbol{\omega}_k\right\| \to 0$ при $k \to \infty$. В свою очередь $\mathbf{z}_k = A^{-1}B\boldsymbol{\omega}_k$, так что

$$\left\|\mathbf{z}_k\right\| \leq \left\|A^{-1}\right\| \cdot \left\|B\right\| \cdot \left\|\boldsymbol{\omega}_k\right\| \to 0$$

Теорема доказана.

### 3.4. Метод простой итерации.

Такое название получил метод, при котором в качестве матрицы $B$ выбирается единичная матрица: $B = E$, а итерационный параметр $\tau$ предполагается независящим от номера итерации $k$. Иными словами, метод простой итерации – это явный стационарный метод, когда очередная итерация $x_{k+1}$ вычисляется по рекуррентной формуле

$$\mathbf{x}_{k+1} = \left(E - \tau A\right)\mathbf{x}_k + \tau\mathbf{f} \quad (97)$$

Будем считать, что матрица $A$ удовлетворяет условию теоремы Самарского, $A = A^* > 0$, тогда формула (87), определяющая границу интервала сходимости по итерационному параметру $\tau$, принимает вид

$$\tau_0 = \inf_{x \neq 0} \frac{2\left(\mathbf{x}, \mathbf{x}\right)}{\left(A\mathbf{x}, \mathbf{x}\right)} = \frac{2}{\sup\limits_{x \neq 0} \dfrac{\left(A\mathbf{x}, \mathbf{x}\right)}{\left(\mathbf{x}, \mathbf{x}\right)}}. \quad (98)$$

Пусть $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$ - ортонормированный базис собственных векторов оператора, соответствующего матрице $A$. В силу положительной определенности все его собственные значения положительны. Будем считать их занумерованными в порядке убывания:

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0 \quad (99)$$

Разложим вектор $\mathbf{x} \neq 0$ по базису собственных векторов

$$\mathbf{x} = \xi_1 e_1 + \xi_2 e_2 + \cdots + \xi_n e_n,$$

тогда

$$\left(\mathbf{x}, \mathbf{x}\right) = \xi_1^{\ 2} + \xi_2^{\ 2} + \cdots + \xi_n^{\ 2},\ \left(A\mathbf{x}, \mathbf{x}\right) = \lambda_1\xi_1^{\ 2} + \lambda_2\xi_2^{\ 2} + \cdots + \lambda_n\xi_n^{\ 2}$$

и

$$\sup_{\mathbf{x} \neq 0} \frac{\left(A\mathbf{x}, \mathbf{x}\right)}{\left(\mathbf{x}, \mathbf{x}\right)} = \sup_{\mathbf{x} \neq 0} \frac{\lambda_1\xi_1^{\ 2} + \lambda_2\xi_2^{\ 2} + \cdots + \lambda_n\xi_n^{\ 2}}{\xi_1^{\ 2} + \xi_2^{\ 2} + \cdots + \xi_n^{\ 2}} = \lambda_1.$$

В результате из формулы (87) следует, что метод простой итерации сходится при любом $\tau$, принадлежащем интервалу

$$0 < \tau < \tau_0 = \frac{2}{\lambda_1}. \quad (100)$$

Дальнейшее исследование метода простой итерации построим на конкретном анализе рекуррентной формулы (97). Введем матрицу оператора перехода

$$S = E - \tau A\,,\; S = S^{*} \tag{101}$$

и перепишем формулу (97) в виде

$$\mathbf{x}_{k+1} = S\mathbf{x}_k + \tau\mathbf{f}\,. \tag{102}$$

При этом погрешность $\mathbf{z}_k = \mathbf{x} - \mathbf{x}_k$ будет удовлетворять аналогичному рекуррентному соотношению, только однородному

$$\mathbf{z}_{k+1} = S\mathbf{z}_k\,. \tag{103}$$

Докажем две леммы, которые позволяют более полно исследовать условия сходимости метода простой итерации.

**Лемма 1**

*Пусть оператор, который порождает матрица $A$, имеет собственный вектор $\mathbf{e}_i$ с собственным значением $\lambda_i$, тогда оператор перехода, который порождается матрицей $S$ (101), также имеет собственный вектор $\mathbf{e}_i$, но с собственным значением*

$$\mu_i(\tau) = 1 - \tau\lambda_i\,. \tag{104}$$

Доказательство элементарно. Оно проводится прямой проверкой

$$S\mathbf{e}_i = (E - \tau A)\mathbf{e}_i = (1 - \tau\lambda_i)\mathbf{e}_i = \mu_i\mathbf{e}_i$$

При самосопряженной матрице $A$ матрица $S$ также является самосопряженной (101). Следовательно, ее норма определяется наибольшим по модулю собственным значением $\mu_i(\tau)$ (104):

$$\|S\| = \max_{1\le i\le n}|\mu_i(\tau)|\,. \tag{105}$$

**Лемма 2**

*Для того, чтобы метод простой итерации сходился к решению системы (62) при любом выборе начального приближения, необходимо и достаточно, чтобы все собственные значения оператора перехода $S$ были по модулю меньше единицы:*

$$|\mu_i(\tau)| < 1,\; 1 \le i \le n \tag{106}$$

*Достаточность*. Условие (106) означает, что норма матрицы $S$, согласно (105), будет меньше единицы: $\|S\| < 1$. В результате получаем

$$\|\mathbf{z}_{k+1}\| \le \|S\|\cdot\|\mathbf{z}_k\| \le \cdots \le \|S\|^k \cdot \|\mathbf{z}_0\| \to 0\,, \text{ при } k \to \infty\,. \tag{107}$$

*Необходимость*. Допустим, что среди собственных значений $\mu_i$ (104) нашлось хотя бы одно $\mu_j$, которое не удовлетворяет условию леммы (106), т. е.

$$|\mu_j| \ge 1.$$

Выберем нулевой член итерационной последовательности в виде $\mathbf{x}_0 = \mathbf{x} + \mathbf{e}_j$, где $\mathbf{x}$ решение системы (62), тогда нулевой член последовательности погрешностей совпадет с собственным вектором $\mathbf{e}_j$ оператора перехода $S$: $\mathbf{z}_0 = \mathbf{e}_j$. В результате рекуррентная формула для следующих членов последовательности погрешностей примет вид:

$$\mathbf{z}_k = S^k\mathbf{e}_j = \mu_j^{\,k}\mathbf{e}_j\,,\; \|\mathbf{z}_k\| = \|\mu_j\|^k \ge 1\,.$$

т. е. $\|\mathbf{z}_k\| \not\to 0$. Необходимость выполнения неравенства (106) для всех собственных значений $\mu_i$ для сходимости метода простой итерации доказана.

Лемма 2 определяет программу дальнейшего исследования сходимости метода простой итерации: нужно установить диапазон изменения параметра $\tau$ при котором все собственные значения удовлетворяют неравенству (106). Это легко сделать. На рис. 1 приведены графики убывающих линейных функций $\mu_i(\tau)$ (104). Все они выходят из одной точки $\tau = 0$, $\mu = 1$ и идут вниз из-за отрицательных коэффициентов при $\tau$, причем быстрее всех убывает функция $\mu_1(\tau)$. Когда она принимает значение $(-1)$, условие (106) для нее перестает выполняться:

$$\mu_1(\tau) = 1 - \tau\lambda_1 = -1, \text{ при } \tau = \tau_0 = 2/\lambda_1 .$$

Найденное значение $\tau_0$ является границей интервала сходимости метода простой итерации

$$0 < \tau < \tau_0 = 2/\lambda_1 . \tag{108}$$

Это неравенство нам уже известно. Оно было получено ранее из теоремы Самарского как достаточное условие сходимости. Дополнительный анализ на основе леммы 2 позволяет уточнить результат. Теперь мы установили, что принадлежность итерационного параметра $\tau$ интервалу (108) является необходимым и достаточным условием сходимости метода простой итерации.

Перейдем к исследованию скорости сходимости метода. Оценка погрешности (107) показывает, что она убывает по закону геометрической прогрессии со знаменателем

$$q(\tau) = \|S\| = \max_{1 \le i \le n} |\mu_i(\tau)|.$$

Рассмотрим рис. 2, который поможет нам провести анализ этой формулы. Он аналогичен рис.1, только на нем приведены графики не функций $\mu_i(\tau)$, а их модулей. При малых $\tau$ все собственные значения $\mu_i(\tau)$ (104) положительны, причем наибольшим из них является $\mu_n(\tau)$, которое убывает с ростом $\tau$ с наименьшей скоростью. Однако с переходом через точку $\tau_0/2$ собственное значение $\mu_1(\tau)$, меняя знак, становится отрицательным. В результате теперь его модуль с увеличением $\tau$ не убывает, а растет и при $\tau \to \tau_0$ приближается к предельному значению – к единице.

Найдем на отрезке $\left[\frac{1}{2}\tau_0, \tau_0\right]$ точку $\tau_*$, в которой убывающая функция $\mu_n(\tau)$ сравнивается с возрастающей функцией $|\mu_1(\tau)| = -\mu_1(\tau)$. Она определяется уравнением

$$\mu_n(\tau) = 1 - \tau\lambda_n = -\mu_1(\tau) = \tau\lambda_1 - 1 ,$$

которое дает

$$\tau_* = \frac{2}{\lambda_1 + \lambda_n} < \tau_0 . \tag{109}$$

В результате получаем:

$$\|S\| = \max_{1 \le i \le n} |\mu_i(\tau)| = \begin{cases} \mu_n(\tau), 0 < \tau \le \tau_* \\ -\mu_1(\tau), \tau_* \le \tau < \tau_0. \end{cases} \qquad (110)$$

Свое наименьшее значение норма матрицы $S$ достигает при $\tau = \tau_*$:

$$\min\|S\| = 1 - \tau_* \lambda_n = \frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n} = \frac{M_A - 1}{M_A + 1}. \qquad (111)$$

Формула (111) показывает, что для плохо обусловленной матрицы даже при оптимальном выборе итерационного параметра $\tau = \tau_*$ норма матрицы $S$ близка к единице, так что сходимость метода простой итерации в этом случае оказывается медленной.

В заключение заметим, что формула (108), определяющая границу интервала сходимости $\tau_0$, и формула (109) для оптимального значения итерационного параметра $\tau_*$ представляют прежде всего теоретический интерес. Обычно при решении СЛАУ наибольшее и наименьшее характеристические числа матрицы $A$ неизвестны, так что подсчитать величины $\tau_0$ и $\tau_*$ заранее невозможно. В результате итерационный параметр $\tau$ нередко приходится подбирать прямо в процессе вычислений методом проб и ошибок.

**Задача 2.**

*Рассмотреть систему двух уравнений с двумя неизвестными*

$$\begin{cases} x_1 + x_2 = 0, \\ x_1 + 2x_2 = 1. \end{cases} \qquad (112)$$

*и построить для нее приближенное решение с помощью метода простой итерации.*

Выпишем сразу решение системы (112)

$$x_1 = -1,\ x_2 = 1, \qquad (113)$$

чтобы потом иметь возможность сравнивать его с членами итерационной последовательности.

Перейдем к решению системы методом простой итерации. Матрица системы имеет вид

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}.$$

Она самосопряженная и положительно определенная, поскольку

$$(A\mathbf{x}, \mathbf{x}) = (x_1 + x_2)x_1 + (x_1 + 2x_2)x_2 = (x_1 + x_2)^2 + x_2^2 > 0.$$

Составим характеристическое уравнение для матрицы $A$ и найдем его корни:

$$\begin{vmatrix} 1-\lambda & 1 \\ 1 & 2-\lambda \end{vmatrix} = \lambda^2 - 3\lambda + 1 = 0,$$

$$\lambda_1 = \frac{3+\sqrt{5}}{2} \approx 2.618,\ \lambda_2 = \frac{3-\sqrt{5}}{2} \approx 0.382$$

С их помощью можно определить границу интервала сходимости $\tau_0$ и оптимальное значение итерационного параметра $\tau_*$:

$$\tau_0 = \frac{2}{\lambda_1} \approx 0.764\,,\ \tau_* = \frac{2}{\lambda_1 + \lambda_2} \approx 0.745\,.$$

Для построения итерационной последовательности выберем какое-нибудь значение итерационного параметра на интервале сходимости, например, $\tau = 1/2$. В этом случае рекуррентная формула для членов итерационной последовательности (102) принимает вид:

$$\mathbf{x}_{k+1} = S\mathbf{x}_k + \frac{1}{2}\mathbf{f}\,,\ \text{где}\ S = \begin{bmatrix} 1/2 & -1/2 \\ -1/2 & 0 \end{bmatrix}$$

Возьмем простейшее начальное приближение $\mathbf{x}_0 = 0$ и выпишем несколько первых членов итерационной последовательности $\mathbf{x}_k$, подсчитывая для каждого из них невязку $\boldsymbol{\psi}_k$ (71). В результате получим:

$$\mathbf{x}_1 = \left\{0, \frac{1}{2}\right\},\ \boldsymbol{\psi}_1 = \left\{\frac{1}{2}, 0\right\},\ \|\boldsymbol{\psi}_1\| = \frac{1}{2},$$

$$\mathbf{x}_2 = \left\{-\frac{1}{4}, \frac{1}{2}\right\},\ \boldsymbol{\psi}_2 = \left\{\frac{1}{4}, -\frac{1}{4}\right\},\ \|\boldsymbol{\psi}_2\| = \frac{1}{2\sqrt{2}},$$

$$\mathbf{x}_3 = \left\{-\frac{3}{8}, \frac{5}{8}\right\},\ \boldsymbol{\psi}_3 = \left\{\frac{1}{4}, -\frac{1}{8}\right\},\ \|\boldsymbol{\psi}_3\| = \frac{\sqrt{5}}{8},$$

$$\mathbf{x}_4 = \left\{-\frac{1}{2}, \frac{11}{16}\right\},\ \boldsymbol{\psi}_4 = \left\{\frac{3}{16}, -\frac{1}{8}\right\},\ \|\boldsymbol{\psi}_4\| = \frac{\sqrt{10}}{16}.$$

Норма невязок, хотя и медленно, но убывает, что говорит о сходимости процесса. Это же видно из сравнения членов итерационной последовательности $\mathbf{x}_k$ с решением системы (113). Медленная сходимость связана с плохой обусловленностью матрицы $A$:

$$M_A = \frac{\lambda_1}{\lambda_2} \approx 6.854\,.$$

### 3.5. Неявные итерационные методы. Метод Зейделя.

Вернемся к общей записи итерационного стационарного процесса в канонической форме (65).

Рассмотрим произвольную квадратную матрицу:

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mm} \end{bmatrix}.$$

Разложим её на сумму трех матриц

$$A = D + T_H + T_B\,, \qquad (114)$$

где $D$ - диагональная часть матрицы $A$, которая содержит элементы $a_{ii}$, стоящие на главной диагонали:

$$D_{ij} = a_{ij}\delta_{ij} = \begin{cases} 0, & i \neq j \\ a_{ii}, & i = j \end{cases},$$

$T_H$ - нижняя треугольная матрица

$$\left(T_H\right)_{ij} = \begin{cases} a_{ij}, & i > j \\ 0, & i \leq j \end{cases},$$

$T_B$ - верхняя треугольная матрица.

$$\left(T_B\right)_{ij} = \begin{cases} 0, & i \geq j \\ a_{ij}, & i < j \end{cases}.$$

В классическом методе Зейделя, записанном в канонической форме, полагают

$$\begin{aligned} &B = D + T_H, \\ &\tau = 1. \end{aligned} \tag{115}$$

В результате формула (65) принимает вид:

$$\left(D + T_H\right)\left(\mathbf{x}_{k+1} - \mathbf{x}_k\right) + A\mathbf{x}_k = \mathbf{f},$$

или

$$\left(D + T_H\right)\mathbf{x}_{k+1} + T_B\mathbf{x}_k = \mathbf{f}. \tag{116}$$

Перейдем от векторной формы записи рекуррентной формулы (116) к построчной:

$$\begin{array}{ll} a_{11}x_1^{k+1} + a_{12}x_2^k + a_{13}x_3^k + \cdots + a_{1n}x_n^k & = f_1 \\ a_{21}x_1^{k+1} + a_{22}x_2^{k+1} + a_{23}x_3^k + \cdots + a_{2n}x_n^k & = f_2 \\ \vdots & \vdots \\ a_{n1}x_1^{k+1} + a_{n2}x_2^{k+1} + a_{n3}x_3^{k+1} + \cdots + a_{nn}x_n^{k+1} & = f_n. \end{array} \tag{117}$$

Уравнения (117) позволяют последовательно рассчитать компоненты вектора $(k+1)$ - ой итерации подобно тому, как это делалось во время обратного хода в методе Гаусса:

$$x_i^{k+1} = \frac{1}{a_{ii}}\left[f_i - \sum_{j=1}^{i-1} a_{ij}x_j^{k+1} - \sum_{j=i+1}^{n} a_{ij}x_j^k\right], \; i = 1,\ldots,n. \tag{118}$$

Формула (118) предполагает, что $a_{ii} \neq 0$, $1 \leq i \leq n$. Если матрица $A$ удовлетворяет условиям теоремы Самарского (84): $A = A^* > 0$, то, согласно неравенству (81), все ее диагональные элементы должны быть строго положительными и, тем самым, не могут обращаться в ноль.

Алгоритм в методе Зейделя прост и удобен для вычислений. Он не требует никаких действий с матрицей $A$. Ранее вычисленные на текущей итерации компоненты $x_j^{k+1}\left(j < i\right)$ сразу же участвуют в расчетах наряду с компонентами $x_j^k\left(j > i\right)$ и, таким образом, не требуют дополнительного резерва памяти, что существенно при решении больших систем.

Сходимость метода Зейделя в случае, когда матрица $A$ удовлетворяет условию теоремы Самарского, т.е. является самосопряженной и положительно определенной, будет доказана в следующем разделе. К этому утверждению добавим без доказательства еще один результат: метод Зейделя сходится для любой системы (62), в которой матрица $A$ обладает свойством диагонального преобладания.

**Задача 3.**

*Рассмотреть систему (112) (см. задачу 2) и построить для нее приближенное решение с помощью метода Зейделя.*

В рассматриваемом случае рекуррентные формулы (118) для построения $(k+1)$ - ой итерации по $k$ -ой итерации принимают вид:

$$\begin{aligned} x_1^{k+1} &= -x_2^k \\ x_2^{k+1} &= \frac{1}{2}\left(1 - x_1^{k+1}\right). \end{aligned} \qquad (119)$$

Принимая, как и при решении задачи 2, за начальное приближение нулевой вектор, подсчитаем по формулам (119) несколько первых итераций, сопровождая этот процесс подсчетом невязки:

$$\mathbf{x}_1 = \left\{0, \frac{1}{2}\right\},\ \boldsymbol{\psi}_1 = \left\{\frac{1}{2}, 0\right\},\ \|\boldsymbol{\psi}_1\| = \frac{1}{2},$$

$$\mathbf{x}_2 = \left\{-\frac{1}{2}, \frac{3}{4}\right\},\ \boldsymbol{\psi}_2 = \left\{\frac{1}{4}, 0\right\},\ \|\boldsymbol{\psi}_2\| = \frac{1}{4},$$

$$\mathbf{x}_3 = \left\{-\frac{3}{4}, \frac{7}{8}\right\},\ \boldsymbol{\psi}_3 = \left\{\frac{1}{8}, 0\right\},\ \|\boldsymbol{\psi}_3\| = \frac{1}{8}.$$

Обсудим полученные результаты. Начнем с невязки. Ее вторая компонента все время остается равной нулю, поскольку второе уравнение системы на каждой итерации выполняется, как видно из (119), точно. Первые компоненты невязки и норма убывают по закону геометрической прогрессии с знаменателем $1/2$, т.е. гораздо быстрее, чем в методе простой итерации. Хорошая сходимость процесса видна также из прямого сравнения членов итерационной последовательности $\mathbf{x}_k$ с точным решением системы $\mathbf{x} = \{-1, 1\}$.

### 3.6. Метод верхней релаксации

Модифицируем метод Зейделя. С этой целью введем параметр $\omega$ и запишем рекуррентное соотношение (65) в виде

$$\left(D + \omega T_H\right)\frac{\left(\mathbf{x}_{k+1} - \mathbf{x}_k\right)}{\omega} + A\mathbf{x}_k = \mathbf{f}\,. \qquad (120)$$

В данном случае

$$B = D + \omega T_H,\ \tau = \omega > 0\,. \qquad (121)$$

При $\omega = 1$ мы возвращаемся к методу Зейделя.

Соотношению (120) можно придать вид

$$\left(\frac{1}{\omega}D + T_H\right)\left(\mathbf{x}_{k+1} - \mathbf{x}_k\right) + A\mathbf{x}_k = \mathbf{f}\,. \qquad (122)$$

Такая форма записи показывает, что параметр $\omega$ влияет на диагональ матрицы $B$.

Для построения алгоритма вычисления очередной итерации нужно разделить в левой части рекуррентной формулы (122) члены, содержащие $\mathbf{x}_k$ и $\mathbf{x}_{k+1}$, и придать ей форму, аналогичную (116):

$$\left(\frac{1}{\omega}D + T_H\right)\mathbf{x}_{k+1} + \left[\left(1-\frac{1}{\omega}\right)D + T_B\right]\mathbf{x}_k = \mathbf{f}\,. \qquad (123)$$

Если перейти от векторной записи к записи типа (117) в виде отдельных уравнений, то можно получить для компонент $x_i^{k+1}$ очередной итерации формулы, структурно похожие на (118):

$$x_i^{k+1} = x_i^k + \frac{\omega}{a_{ii}}\left(f_i - \sum_{j=1}^{i-1} a_{ij}x_j^{k+1} - \sum_{j=i}^{n} a_{ij}x_j^k\right),\ i = 1,\ldots,n\,. \qquad (124)$$

Исследуем условия сходимости метода верхней релаксации при дополнительном предположении, что матрица $A$ удовлетворяет условиям теоремы Самарского (84). Самосопряженность матрицы $A$ означает, что $T_H^* = T_B$, $T_B^* = T_H$. Отсюда следует

$$\left(T_H\mathbf{x},\mathbf{x}\right) = \left(T_H^*\mathbf{x},\mathbf{x}\right) = \left(T_B\mathbf{x},\mathbf{x}\right). \qquad (125)$$

Составим для рассматриваемого случая матрицу $B - \frac{\tau}{2}A$. Согласно (121)

$$B - \frac{\tau}{2}A = \left(D + \omega T_H\right) - \frac{\omega}{2}\left(D + T_H + T_B\right) = \left(1 - \frac{\omega}{2}\right)D + \frac{\omega}{2}\left(T_H - T_B\right). \qquad (126)$$

Запишем условие ее положительной определенности

$$\left(\left(B - \frac{\tau}{2}A\right)\mathbf{x},\mathbf{x}\right) = \left(1 - \frac{\omega}{2}\right)\left(D\mathbf{x},\mathbf{x}\right) > 0\,. \qquad (127)$$

Второе слагаемое в выражении (126) не дает вклада в квадратичную форму (127) в силу соотношения (125).

Матрица $A$ является, по предположению, положительно определенной. Следовательно, все ее диагональные элементы строго положительны: $a_{ii} > 0$, $1 \le i \le n$. Это означает положительную определенность матрицы $D$: $\left(D\mathbf{x},\mathbf{x}\right) > 0$. В результате знак выражения (127) определяется знаком первого множителя, так что достаточное условие для сходимости итерационной последовательности метода верхней релаксации принимает вид:

$$0 < \omega < 2 \qquad (128)$$

Метод Зейделя, соответствующий случаю $\omega = 1$, удовлетворяет этому условию.

Можно поставить вопрос об оптимальном выборе параметра $\omega = \omega_*$, при котором метод сходится быстрее всего. Теоретическое исследование, на котором мы не будем останавливаться, показывает, что такое значение существует и может быть выражено через наибольшее и наименьшее собственные значение матрицы $A$. Однако на практике его приходится подбирать экспериментально методом проб и ошибок, поскольку найти $\lambda_{\min}$ и $\lambda_{\max}$ с достаточной точностью удается в редких случаях.

**Задача 4**

*Построить приближенное решение системы (112) методом верхней релаксации, полагая* $\omega = 4/3$.

Выпишем для рассматриваемого случая матрицы $\frac{1}{\omega}D + T_H$ и $\left(1 - \frac{1}{\omega}\right)D + T_B$, определяющие итерационный процесс:

$$\frac{3}{4}D + T_H = \begin{bmatrix} 3/4 & 0 \\ 1 & 3/2 \end{bmatrix}, \ \frac{1}{4}D + T_B = \begin{bmatrix} 1/4 & 1 \\ 0 & 1/2 \end{bmatrix}.$$

С их помощью рекуррентное соотношение (123), записанное покомпонентно, принимает вид:

$$\frac{3}{4}x_1^{k+1} + \frac{1}{4}x_1^k + x_2^k = 0,$$

$$x_1^{k+1} + \frac{3}{2}x_2^{k+1} + \frac{1}{2}x_2^k = 1.$$

Выражая из первого соотношения $x_1^{k+1}$, из второго $x_2^{k+1}$, получим окончательные расчетные формулы для компонент очередной итерации:

$$x_1^{k+1} = -\frac{1}{3}x_1^k - \frac{4}{3}x_2^k,$$

$$x_2^{k+1} = \frac{2}{3} - \frac{2}{3}x_1^{k+1} - \frac{1}{3}x_2^k.$$

Примем, как и в предыдущих случаях, за начальное приближение нулевой вектор и сделаем три итерации. При этом для каждой из них подсчитаем невязку (71), позволяющую следить за сходимостью процесса

$$\mathbf{x}_1 = \left\{0, \frac{2}{3}\right\}, \ \boldsymbol{\psi}_1 = \left\{\frac{2}{3}, \frac{1}{3}\right\}, \ \|\boldsymbol{\psi}_1\| = \frac{\sqrt{5}}{3} \approx 0.745,$$

$$\mathbf{x}_2 = \left\{-\frac{8}{9}, \frac{28}{27}\right\}, \ \boldsymbol{\psi}_2 = \left\{\frac{4}{27}, \frac{5}{27}\right\}, \ \|\boldsymbol{\psi}_2\| = \frac{\sqrt{41}}{27} \approx 0.237,$$

$$\mathbf{x}_3 = \left\{-\frac{88}{81}, \frac{256}{243}\right\}, \ \boldsymbol{\psi}_3 = \left\{-\frac{8}{243}, \frac{5}{243}\right\}, \ \|\boldsymbol{\psi}_3\| = \frac{\sqrt{89}}{243} \approx 0.039.$$

Поведение невязок, а также сравнение членов итерационной последовательности $\mathbf{x}_k$ с точным решением системы $\mathbf{x} = \{-1, 1\}$ показывают сходимость процесса, более быструю, чем в методе Зейделя. Выбранное значение параметра $\omega = 4/3$ оказалось близким к оптимальному $\omega = \omega_*$.